ALTERNATIVE MODELS FOR CPS INCOME IMPUTATION

Roderick J.A. Little and Michael E. Samuhel, ASA/Bureau of the Census

1. INTRODUCTION

In recent years, a variety of methods have been developed for carrying out statistical analyses for data sets with missing values. For reviews of the literature, see Afifi and Elashoff (1966), Hartley and Hocking (1971), Orchard and Woodbury (1972), Dempster, Laird and Rubin (1977), Little (1982). A weakness in this literature is the absence of comparisons between methods in realistic applied settings.

Comparisons on real, incomplete data sets are rare, since in such situations truly objective measures of the utility of missing data methods require estimates from the hypothetical complete data with the true values of missing data filled in. These values are rarely available, since if they were the data would no longer be incomplete. In the Income Supplement of the Current Population Survey (CPS) questions concerning annual earnings are not answered by all individuals in the survey. The CPS data are later matched to IRS records in the 1973 CPS - Administrative Record Exact Match File (Aziz, Kilss and Scheuren, 1978) yielding IRS earnings values for respondents and nonrespondents to the CPS earnings questions. The IRS and CPS values cannot be equated because of matching errors and intrinsic differences in the way the two variables are reported. However, the IRS earnings values do provide valuable information for comparing alternative imputation methods. In particular, Greenlees, Reece, and Zieschang (1982) use the nonresponse pattern in the CPS to delete IRS values, and then compare imputations for the deleted values with the true values. The key advantage over Monte Carlo methods is that the IRS values are deleted according to a realistic data mechanism.

This paper and a companion article (David and Triest, 1983) report research on another IRS/CPS matched file, with data from the March 1981 Income Supplement to the CPS matched to 1980 IRS Income records. The objective is to evaluate the current Census Bureau methodology for imputing earnings data, and to compare it with alternatives. This paper discusses the CPS Hot Deck (HD) procedure and a variety of alternative procedures based on regression models for wage and salary amounts. The paper by David and Triest presents the results of empirical comparisons between these methods based on the CPS/IRS matched data.

In the next section we provide a brief description of the CPS hot deck procedure, discuss its theoretical strengths and limitations, and introduce the alternative regression based methods. In Sections 3 and 4, two regression models for imputing wages and salary (WS) amounts are presented, one with the logarithm of wages and salary amounts as the dependent variable, and the other with the wage rate as dependent variable. In section 5 we review methods for adding residuals to the predictors from these models and propose a new variant of these methods with attractive properties. In section 6 some initial numerical results are presented for the methods - see David and Triest (1983).

2. THE CPS HOT DECK AND ALTERNATIVES

Census methods for imputing (or in Census terminology, allocating) missing income items have developed continuously over the last 20 years. In the early 1960's, individuals who refused to report their income were simply ignored in published data. The losses involved in this procedure are not serious given relatively low nonresponse rates. The real problem is the bias introduced in estimates because respondents are not a random subsample of the sampled individuals. Beginning with the 1962 CPS, a hot deck procedure assigned the income of a matched individual to each person who did not report his income. The method has been refined since 1962 by increasing the number of variables used to define a match and by modifying the treatment of multiple income items to help preserve their covariance structure. A historical survey of the CPS hot deck and an evaluation of its impact on the variance of survey estimates is provided by Oh and Scheuren (1980).

The current imputation scheme for earnings (Welniak and Coder, 1980), initially classifies nonrespondents into one of eight groups, according to the combination of missing values for earnings recipiency and amount, work experience and longest job. Within each group, nonrespondents are matched with respondents with similar values of the earnings, work experience and job information available for that group, and other covariates, namely sex, age, race, educational attainment, relationship to family head, marital status, number of children, type of residence, region of residence, and other income recipiency pattern. The nonrespondent is then assigned the matched respondent's values of the missing items.

The number of variables used to define matches is extensive. In practice, for many nonrespondents no respondents can be found which match on all these items. In such cases matches are found at a lower level of detail, by omitting some matching variables and reducing the number of categories in others. At the lowest level of detail in the group one, only sex, age in three categories, race in two categories education in three categories relationship to head in two categories weeks worked last year in 2 categories, full time/part time status, 46 category occupation-industry coding, class of worker in 2 categories and earnings recipiency pattern are used to define the match. Not all the nonrespondents in group one are matched at the least detailed level in group one. Non-matched individuals drop to a group with less information for matching. A few individuals drop down to the lowest level in group eight, where only sex, age in four categories and education in three categories are used to determine matches. Matches are found for all nonrespondents at or before this group and level, but the quality of the match varies considerably, depending on the availability of suitable respondents for matching.

Lillard, Smith and Welch (1982) provide a penetrating discussion of the theoretical properties of the HD scheme. They note that for a given level of detail, the hot deck is similar to fitting a fully interactive analysis of variance and then imputing for nonrespondents the predicted mean plus an empirically based residual. For example, suppose three categorical covariates X_1, X_2 and X_3 are used to define a match, and let $y_{ijk\ell}$ be the earnings for respondent ℓ in a cell with X_1=i, X_2=j and X_3=k. A fully interactive model specifies

$$y_{ijk\ell} = \mu_{ijk} + \varepsilon_{ijk\ell}, \qquad (1)$$

where μ_{ijk} is the expected earnings in the cell and $\varepsilon_{ijk\ell}$ is a random error. Suppose a nonrespondent in this cell is matched to respondent m. The hot deck imputed value y_{ijkm} can be decomposed as

$$y_{ijkm} = \bar{y}_{ijk} + r_{ijkm},$$

where $\bar{y}_{ijk}$ is the predicted mean value from estimating μ_{ijk} and

$$r_{ijkm} = (y_{ijkm} - \bar{y}_{ijk})$$

is the residual for a respondent chosen randomly from the (i,j,k)th cell.

Since the model is fully interactive and interval scaled variables such as age are grouped into categories, the method is relatively agnostic about the appropriate form of the equation relating earnings to the predictors. It does however make the important (and often unjustified) assumption that respondents and nonrespondents have the same earnings distribution within the cell defined by matching predictors. That is, the nonresponse mechanism is assumed ignorable, in the sense discussed by Rubin (1976). Furthermore, the precision of the imputation may be compromised by omitting detail from the model. For example, Lillard, Smith and Welch observe that the mean imputed income for nonreporting white male lawyers in the 1980 CPS is $33,448 when detailed occupational coding is used in the match, and only $15,594 when one digit coding only used. Nonrespondents from rare population subclasses with particularly high or low incomes tend to be difficult to match, and as a result are pulled towards the mean of the income distribution by the lack of detail at the level a match is made.

A second shortcoming of the HD scheme is that a donor for imputed values may be used more than once. This results in particular respondents effectively receiving abnormally large weights, a procedure which reduces nonresponse bias at the expense of increased variance of estimates in repeated sampling (see, for example, Scheuren, 1983). The problem of multiple donors may be important, given that the full detail in level 1 imputations implies a matrix of

$$2 \times 5 \times 3 \times 5 \times 5 \times 3 \times 4 \times 2 \times 375 \times 5 \times 3 \times 4 \times 7 = 2.83 \times 10^9$$

cells while the 130,000 cases of CPS data include about 17,000 nonreporters of income items.

The problems of high dimensionality appear even more relevant when omissions from the hot deck algorithm are considered. Nonresponse stems from both random events and inhibitions affecting reporting. Respondents who regard the divulging of income information as inappropriate may refuse in the extreme, but they may also lie and distort in order to evade reporting. Some indications of this type of behavior can be found in the reporting of other items, in the rounding of reported data, and in aspects of the interviewing situation, such as the presence or absence of others (Cannell and Henson, 1974; Lansing, Ginsberg and Braten, 1961). Such information ought also to be incorporated in the model (1). However, the purpose of such variables will be to correct for response bias, so that imputed values are closer to population values. Incorporating additional variables in the HD is clearly difficult, as more and more situations arise in which lower-level matrices are used for matching.

One way of thinking about the HD is that it selects values from a nested family of models, such as (1), in which the selection is determined by the drawing of particular individuals into the sample. This makes the imputation outcome a random function of both the sampling mechanism and the real variation in the population.

A natural alternative to the hot deck is to base inferences on a more parsimonious model for earnings. For example, setting interactions in (1) equal to zero yields

$$y_{ijk\ell} = \mu + \alpha_{1i} + \alpha_{2j} + \alpha_{3k} + \varepsilon_{ijk\ell}, \qquad (2)$$

where the parameters $\{\alpha_{1i}\}$, $\{\alpha_{2j}\}$ and $\{\alpha_{3k}\}$ are identified by suitable linear constraints, for example $\Sigma\alpha_{1i} = \Sigma\alpha_{2j} = \Sigma\alpha_{3k} = 0$.

Imputed values for nonrespondents in cell (i,j,k) may be the predicted means from (2),

$$\hat{y}_{ijk} = \hat{\mu} + \hat{\alpha}_{1i} + \hat{\alpha}_{2j} + \hat{\alpha}_{3k}, \qquad (3)$$

where $\hat{\mu}$, $\hat{\alpha}_{1i}$, $\hat{\alpha}_{2j}$, and $\hat{\alpha}_{3k}$ are estimates of the parameters. This procedure is a special case of regression imputation. An alternative procedure, which better preserves the distribution of earnings in the imputed data, imputes values of the form

$$y_{ijk} = \hat{y}_{ijk} + r_{i'j'k'\ell'} \qquad (4)$$

where

$$r_{i'j'k'\ell} = y_{i'j'k'\ell} - \hat{\mu} - \hat{\alpha}_{1i'} - \hat{\alpha}_{2j'} - \hat{\alpha}_{3k'}$$

is a randomly selected residual. Note that the selected residual need not be in the same cell as the nonrespondent. This compromise between hot deck and regression imputation is attributed to Fritz Scheuren in Schieber (1978). Since the residual is not restricted to the nonrespondent's cell, the method works when there are no respondents in that cell, unlike the CPS hot deck method. The model is assumed to capture the relationship between earnings and X_1, X_2, and X_3, so that the residuals have no structure.

The imputations (3) and (4) will improve on the CPS hot deck, if the model assumptions are valid, since the variance added by the imputations will be smaller. To take an extreme case, a cell with one respondent and five nonrespondents uses the respondent value five times in the CPS hot deck, yielding estimates with inflated standard errors. This property is avoided in the model-based method, where residuals are selected from the whole respondent file. A more potent argument for the modelling procedure is that a large number of covariates can be included simultaneously in the model, with greatly reduced restrictions on the level of detail compared with those imposed in the CPS hot deck. This allows potentially more accurate predictions of the respondent means, and weakens the ignorability assumption about the nonresponse mechanism.

Groups 2 to 8 of nonrespondents have other items besides earnings missing, and require multivariate imputation models. In this paper we concentrate on imputing wages and salary (WS) amounts to individuals where the earnings amounts alone are missing. Regression models like (2) are fitted to the respondent data. Categorical variables are represented by indicator variables and interval scaled variables such as age and education are treated as continuous covariates.

3. MODELLING THE LOGARITHM OF WAGES AND SALARY

3.1 Introduction

To gain some experience in modelling wages and salary, we selected a systematic ten percent sample of the 1980 CPS Income Supplement, yielding 13,831 individuals. The models for wages and salary amounts were fitted on a subset of 7037 of these individuals, consisting of income recipients with values of all the variables used to predict earnings present and with wages and salary amounts reported of more than $100. The resulting WS equations were used to predict the missing WS values in the CPS. An important assumption underlying this approach is that the regression relationships estimated from the respondent sample also apply to the nonrespondents. In other words, the response mechanism is assumed to be ignorable, in the sense discussed by Rubin (1976).

In developing our model we build on previous earnings models described by Lillard and Willis (1978), Greenlees, Reece and Zieschang (1982) and Betson and Van der Gaag (1983). Greenlees, Reece and Zieschang fitted an additive model relating log (WS) to linear and quadratic terms in education and work experience, race (white vs others), urbanity, region, and one digit occupational codes. The sample base was highly restricted, including heads of primary families in which the head was at least 14 years old, was married with spouse present, had a nonfarm residence, had no farm or self-employment income, was employed full time for the full year in the private nonagricultural sector and filed a joint tax return. The authors imposed these restrictions to reduce computation burden for an alternative estimation method based on a stochastic censoring model. The restricted sample base limits the utility of the model for CPS imputations. In particular, the restrictions eliminate about 80% of the sample individuals we wish to model.

Lillard and Willis (1978) modelled 1967-1973 earnings data from the University of Michigan Panel Study of Income Dynamics. Their sample base consisted of 1144 male heads of household aged 18-58 in 1967 who were not disabled, retired or a full time student during the period and who reported positive annual hours and earnings each year. The most detailed version of their model included variables similar to those in Greenless, Reece and Zieschang (1982), plus marital status, distance, union/nonunion, unemployment, and variables relevant to their longitudinal analysis which need not concern us here. They also included various two way interactions between schooling, work experience, race and unemployment.

A more detailed model is required for our data base. Our model includes females, is not restricted to household heads, and includes individuals with part time or self-employment income. The modelling of WS amounts for part time or intermittent workers is particularly important, since the logarithmic transformation of the dependent variable makes individuals with low annual WS amounts influential in the regression. A natural way of modelling these wage amounts is by including log (weeks worked), (LNWK), and log (hours per week worked), (LNHR), as covariates in the regression.

After some experimentation, the five variable quadratic surface,

$$\mathrm{LNWK}, \mathrm{LNHR}, \mathrm{LNWK}^2, \mathrm{LNHR}^2, \mathrm{LNWK}*\mathrm{LNHR}$$

was used to capture the effects of these factors on log (WS). The interaction and squared terms proved to be highly significant when added to the regression. This finding is at variance with existing models of earnings, as in Lillard-Willis (1978) and Betson-van der Gaag (1983). Details of the model are available from the authors upon request.

3.2 Estimating WS from the model

Writing y for log (WS), our predictions are based on the model

$$y = \beta' x + \varepsilon,$$

where x is the set of predictors and ε is normal with mean 0, variance σ^2. Exponentiating and taking expectations with respect to distribution of ε yields

$$E[\exp(y)]=\exp(\beta x' + \sigma^2/2), \qquad (5)$$

by properties of the lognormal distribution. The prediction of y from the model takes the form

$$\hat{y} = \hat{\beta}' x = \beta' x + \eta$$

where $\hat{\beta}$ is the least squares estimate of β, and η is a prediction error. The latter is normal with mean 0, and variance $x'(X'X)^{-1}x\ \sigma^2$ under the model, where X is the design matrix of the respondent sample. Hence,

$$E[\exp(\hat{y})]=\exp(\beta x' + x'(X'X)^{-1} x \sigma^2/2) \qquad (6)$$

Comparing (5) and (6), unbiased predictions of WS are obtained from the equation

$$WS=\exp(\hat{y}+(\sigma^2/2)(1-x'(X'X)^{-1}x)). \qquad (7)$$

In implementing (7), σ^2 is unknown and is replaced by the residual mean square s^2 from the regression. The resulting correction is a refinement of the usual adjustment

$$\hat{WS}^* = \exp(\hat{y}+s^2/2), \quad (8)$$

which ignores the sampling variation of the estimated regression coefficients. The ratio WS/WS^* is order (1/n), where n is the sample size. Thus, for our data (7) and (8) are approximately equal, since the sample size is large. We used (7) to generate predictions, since the order (1/n) refinement of (8) is readily available from standard regression output.

For our model $s^2=0.2786$ and the effect of (7) is to increase the predictions $e^{\hat{y}}$ by approximately $100(\exp(s^2/2)-1) = 15\%$, a non-trivial adjustment.

One modification of (7) which proved worthwhile was an adjustment for heteroscedasticity. A table of the mean squared residual (MSR) against the predicted log WS value ($\hat{y}$)suggested a downward linear trend. Unweighted linear regression yielded the expression

$$MSR = 0.2786 - .068 (y - 8.87). \quad (9)$$

This expression was substituted for s^2 in equation (7), and yielded an average predicted WS value for respondents which closely matched the observed mean. The expression (9) could also have been used to define weights for the linear regression, but this refinement was not thought likely to have much impact on the predictions and hence was omitted.

4. MODELLING THE WAGE RATE

Instead of modelling the logarithm of wages and salaries and using the reported weeks and hours as independent variables, an alternative is to model the wage rate per hour, RATE = WS/WKHR, where WKHR is the product of weeks worked and hours worked per week. The model of the wage rate will hereafter be referred to as the ratio model. The ratio model uses the same set of independent variables as the log model, except that the main effects of weeks worked and hours worked replaced the five variable quadratic surface for these variables discussed in the log model. A more parsimonious representation appears justified, given the partial adjustment inherent in the denominator (WKHR) of the dependent variable. The ratio model was fitted by weighted regression with weights defined as

$$WT= WKHR/1664.43,$$

the normalizing constant 1664.43 being the average value of WKHR for cases in the regression. Details of the model are available from the authors upon request.

The main reason for adopting the ratio model rather than the log model is that the WS predictors are a linear transformation of the predictions from the model, found by multiplying the predicted wage rate by the appropriate value of WKHR. Thus the problems of untying the log transformation for the log model, discussed in 3.2, are avoided. In particular, by the properties of linear regression, the observed and predicted mean WS amounts automatically match in subclasses of the sample defined by categorical regressors in the model. A related advantage of the rate model is that it avoids the excessive weighting of low income observations implicit in the log model, which tends to distort the relative importance of the predictor variables. For example, occupational group has a much greater predictive power in the ratio model then in the log model, in which weeks worked and hours worked dominate all other effects.

The ratio model also has limitations. Inaccuracies in the reporting of weeks worked and hours worked in the CPS may distort the dependent variable. Also the log model always yields positive predicted WS amounts, whereas the ratio model predicted a small number of negative WS amounts when applied to the CPS data. These values were set to zero in comparisons. Results reported in David and Triest (1983) indicate negligible differences between the models in their ability to predict IRS wages and salary amounts.

5. ADDING RESIDUALS TO THE MODEL PREDICTIONS

The models discussed in sections 3 and 4 impute for missing WS amounts a mean of the predictive distribution, conditional on the included predictors. As a result, the distribution of the imputed values has a smaller variance than the distribution of the true values, even if the assumptions of the model are valid. One strategy for adjusting for this attenuation is to add random errors to the predictive means. These errors can take the form of random normal deviates or randomly selected residuals from the model.

For the log model in section 3, a simple procedure is to impute

$$\hat{WS}_1 = \exp(\hat{y} + Zs), \quad (10)$$

where $\hat{y}$is the predicted log (WS) amount, Z is a standard normal deviate and s is the residual standard deviation. This method performed poorly in preliminery comparisons, mainly because too much noise is added to the large values of PRED, producing some extremely large WS predictors after exponentiation. Improved predictions are obtained from

$$W\hat{S}_2 = \exp(\hat{y}+ MSR*Z), \quad (11)$$

where MSR is calculated from equation (9), which reflects the decline in the residual variance as PRED increases.

Both (10) and (11) suffer from a reliance on the assumption of normal residuals from the model. A more flexible approach is to impute

$$WS=\exp(\hat{y}+\hat{r}_j), \quad (12)$$

where r_j is the residual from a randomly selected respondent j. Previous authors (eg.,Kalton and Kish, 1981) have suggested selecting r_j by some form of random sampling (simple random with or without replacement, or stratified by the residual values) applied to the whole respondent file. We propose the preliminary step of classifying respondents and nonrespondents according to their values of y, and then assigning residuals to nonrespondents from respondents in the same cell. For the log model, cells were formed by intervals of exp ($\hat{y}$) of width \$2000.

This correction for attenuation has the appealing property that, when applied to respondent data, the distribution of WS values is approximately preserved. Suppose that within a cell with a fixed value of $\hat{y}$, donor residuals are assigned to nonrespondents (recipients) by simple random sampling without replacement. If the respondents are treated as both donors and recipients, the number of donors and recipients are equal, and the method effectively permutes the WS values in the cell. To see this, note that the predicted WS value for unit i is

$$\hat{WS}_i = \exp\ [\hat{y}_i + y_{\pi(i)} - \hat{y}_{\pi(i)}],$$

where $\pi(.)$ denotes a permutation of the respondents in the cell. If y_i is constant within the cell, then

$$\hat{y}_{\pi(i)} = \hat{y}_i \text{ and}$$

$$\hat{WS}_i = \exp(y_{\pi(i)}) = WS_{\pi(i)},$$

the observed WS value for respondent $\pi(i)$. That is,the method permutes the wages and salary values in the cell. Hence, the distribution of the respondent WS values is unchanged by the procedure. In practice, the stratification by exp $(\hat{y})$ into \$2000 intervals results in some variability in $\hat{y}$ within strata, so the distribution is only approximately preserved. Note that a method which does not stratify on $\hat{y}$ would not produce a permutation of the WS values when applied to the respondents, and hence may not preserve the distributional properties of the respondent sample.

Similar adjustments can be made to the ratio model. The addition of a normal deviate with the appropriate variance is not advisable, since the distribution of residuals, classified by the predicted wage rate, shows considerable skewness, particularly for large and small predicted rates. In particular, the method can produce large negative estimates of wages and salary. The method chosen for comparisons calculates imputations of the form

$$\hat{WS}_4 = (RATE + r_j)WKHR, \tag{13}$$

where r_j is a residual selected within strata formed by predicted WS values. In both (12) and (13), donors are selected for recipients by systematic sampling of the respondents' residuals, ordered from lowest to highest value. This method of selection minimizes the added variance of estimates caused by the appended residuals (Kalton and Kish, 1981).

6. RESULTS OF THE MODEL FITTING

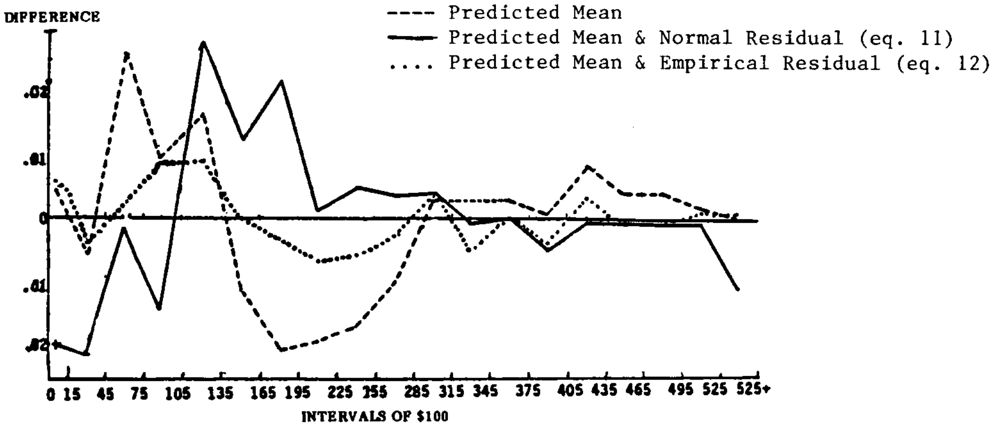
The log model discussed in section 3 had an R-squared of 0.813 and a residual standard deviation of s=0.53. The R-squared value is sensitive to the way in which low WS values are treated, and is largely determined by the modelling of weeks worked and hours worked. The residual standard deviation is somewhat larger than the value s=0.40 obtained by Greenlees, Reece and Zieschang (1982), but their model was fitted to a much more homogenous population, so this result is not surprising. A regression of the logarithm of the residual mean square from the log model on a restricted set of covariates revealed that the model predicts noticeably better for females than for males, and much better for individuals for whom wages and salary is the only source of income than for individuals with additional self-employment income. The latter result conforms with expectations, since the weeks worked and hours Worked variables apply to all sources of income, and thus are not reliable predictors for WS when more than one source is involved.

The ratio model discussed in section 4 had an R-squared of 0.41 and a residual standard deviation s=3.75, which can be compared with the mean wage rate of 7.15 for the respondent sample. The R-squareds for the log and ratio model are of course not comparable, since the dependent variables are quite different.

The distributional properties of imputations based on these models are examined in Figure 1. The solid line shows the deviations of the model predictions from the respondents values. These deviations are negative for the extremes of the range and positive in the middle, indicating the attenuation of the distribution of conditional means noted in section 4.

A prediction method which perfectly reflects the distribution of observed WS values should produce a horizontal line through the origin in Figure 1. The results of adding errors to the predicted means from equations (11) and (12) are shown in the dashed and

Figure 1. The Deviations of Log Model Estimates for Wages and Salaries from the Respondent Values.



dotted lines in Figure 1, respectively. Both methods remove the attenuation of the distribution at the high and low values, but the empirical residual method (equation (11)) is far better at reproducing the observed distribution of WS amounts, basically because the normality assumption underlying equation (11) is not justified. This plot confirms the superiority of the empirical residual method applied within strata formed by the predicted values, which was predicted by the theoretical argument in section 4.

We have shown in this paper how relatively detailed models for imputation of wages and salary amounts can be constructed, and how empirical residuals can be added to the imputations from these models to yield imputations with excellent distributional properties. The acid test of these methods is how well they actually predict the missing WS items. This is the subject of the companion paper by David and Triest (1983), to which we refer the interested reader.

ACKNOWLEDGEMENTS

Roderick J.A. Little is Associate Professor of Biomathematics, UCLA School of Medicine, and Michael E. Samuhel is a member of the Statistical Research Division, Bureau of the Census. This paper reports research conducted by the authors with Martin David and Robert Triest under the American Statistical Association / Census Bureau Fellowship program, supported by grants from the National Science Foundation and the U.S. Bureau of the Census. The authors gratefully acknowledge the advice and guidance of John Coder, Gordon Green and Chuck Nelson of the Population Division of the Census Bureau.

REFERENCES

Afifi, A.A. and Elashoff, R.M. (1966), "Missing observations in multvariate statistics I: Review of the literature ," Journal of the American Statistical Association, 61, 595-604

Aziz, F., Kilss, B. and Scheuren, F.(1978), 1973 Current Population Survey - Administrative Record Exact Match File Codebook , Part I - Code Counts and Item Definitions, Washington, D.C., U.S. Department of Health, Education and Welfare.

Betson, D. and Van der Gaag, J. (1983), "Working married women and their impact on the distribution of welfare in the United States," Working paper, Institute for Research on Poverty, University of Wisconsin.

Cannel, C. and Henson, R. (1974), "Incentives, motives and response biase," Annals of Economic and Social Measurement.

David, M.R. and Triest, R.K. (1983), "The CPS hot deck: an evaluation using IRS records", American Statistical Association, 1983 Proceedings of the Section on Survey Research Methods.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977), Maximum likelihood from incomplete data via the EM algorithm, "Journal of the Royal Statistical Society, Series B, 39, 1-38 (with discussion).

Greenless, W.S., Reece, J.S. and Zieschang, K.D. (1982) "Imputation of missing values when the probability of response depends on the value being imputed," Journal of the American Statistical Association, 77, 251-261.

Hartley, H.O. and Hocking, R.R. (1971), "The analysis of incomplete data," Biometrics,27, 783-808.

Kalton, G. and Kish, L. (1981), "Two Efficient Random Imputation Procedures," American Statistical Association, 1981 Proceedings of the Section on Survey Research Methods, 146-151.

Lansing, J.B., Ginsberg, G. and Braten, K. (1961), An Investigation of Response Error, University of Illinois Press.

Lillard, L.A., and Willis, R.J. (1978), "Dynamic Aspects of Earnings Mobility," Econometrics, 46, 985-1011.

Lillard, L., Smith, J.P. and Welch,(1982)"What do we really know about wages: The importance of non-reporting and Census imputation," The Rand Corporation, 1700 Main St., Santa Monica, CA 90406.

Little, R.J.A. (1982), "Models for nonresponse in sample surveys," Journal of the American Statistical Association, 77, 237-250.

Oh, H.L. and Scheuren, F. (1983), "Weighting adjustments for unit nonresponse," to appear in Incomplete Data: The Theory of Current Practice, National Academy of Sciences, (In press).

Oh, H.L. and Scheuren, F. (1980), "Estimating the variance impact of missing CPS income data," American Statistical Association, 1980 Proceedings of the Section on Survey Research Methods, 408-415.

Orchard, T. and Woodbury, M.A. (1972), "A missing information principle: theory and applications," Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability, 1, 697-715.

Rubin, D.B. (1976), "Inference and missing data," Biometrika,63, 581-592.

Schieber, S. J. (1978), "A comparison of three alternative techniques for allocating unreported social security income on the Survey of the Low-Income Aged and Disabled," American Statistical Association, 1978 Proceedings of the Section on Survey Research Methods, 212-218.

Welniak, E.J. and Colder, J.E. (1980), "A measure of the bias in the March CPS earnings imputation scheme," American Statistical Association, 1980 Proceedings of the Section on Survey Research Methods, 421-425.